*Research Article*

Research and Politics
October-December 2015: 1–6

DOI: 10.1177/2053168015606749
rap.sagepub.com
SAGE

# Political accountability and autonomous weapons

**James Igoe Walsh**

**Abstract**

Autonomous weapons would have the capacity to select and attack targets without direct human input. One important objection to the introduction of such weapons is that they will make it more difficult to identify and hold accountable those responsible for undesirable outcomes such as mission failures and civilian casualties. I hypothesize that individuals can modify their attribution of responsibility in predicable ways to accommodate this new technology. The results of a survey experiment are consistent with this; subjects continue to find responsible and hold accountable political and military leaders when autonomous weapons are used, but also attribute responsibility to the designers and programmers of such weapons.

**Keywords**

Autonomous weapons, accountability, responsibility attribution

## Introduction

Machines equipped with artificial intelligence now make decisions without direct human input. Examples include computers that play chess, decide what news, products, or status updates Internet users see, and that assist in medical diagnosis. Soon artificial intelligence may control machines such as automobiles, aircraft, power stations, and weapons. Decisions or malfunctions by such computers could have lethal consequences. Who takes the blame when an autonomous device kills or injures a human being? Answers to this question have important political consequences. Effective democratic governance requires that citizens be able to identify and hold accountable those who make decisions that produce undesired outcomes (see, among many, McGraw, 1990; Peffley, 1984; Powell and Whitten, 1993; Rudolph, 2003). But allowing machines to make lethal decisions could create a 'responsibility gap' (Matthias, 2004), as it is not clear how such machines could be held accountable.

After defining autonomous weapons, I develop hypotheses about the effects they might have on the attribution of responsibility and accountability, and then assess these with a survey experiment. Drawing on the work of Schulzke (2013) and empirical studies of responsibility attribution to multiple actors, I hypothesize that leaders would still be held accountable, as they are seen as responsible for the decision to employ autonomous weapons. Furthermore, individuals adjust their attributions of responsibility to include a wider and more complex range of actors, such as those involved in the design and programming of autonomous weapons. The results of the survey experiment are consistent with these hypotheses. This suggests that existing mechanisms for attributing responsibility and holding actors accountable, such as the military chain of command, could with some adaptations play the same role in situations where autonomous weapons are utilized. The conclusion briefly discusses the possible influence of public attitudes on responsibility attribution should autonomous weapons ever be developed.

## What are autonomous weapons?

The United States Department of Defense defines autonomous weapons as 'a weapons system that, once activated,

University of North Carolina at Charlotte, USA

**Corresponding author:**
James Igoe Walsh, Department of Political Science, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte NC 28223 USA.
Email: jwalsh@uncc.edu @jamesigoewalsh

can select and engage targets without further intervention by a human operator' (Department of Defense, 2012: 13; similar definitions are used by Human Rights Watch, 2015, and the rapporteur of the UN Human Rights Council, 2013; for a discussion, see Wagner, 2014). This definition 'includes human-supervised autonomous weapons systems that are designed to allow human operators to override operation of the weapon system'. Such human-supervised weapons already exist. Examples include missile defense systems, where humans position, activate, and supervise the weapon, but rely on it to follow predetermined rules to identify and attack incoming missiles more quickly and precisely than a human operator could do so. Here humans are 'on the loop'; they decide when and where to deploy the weapon, and can intervene to prevent its operation. For this reason, such weapons might be better described as semi-autonomous, as they lack the capacity to engage in combat without direct and immediate human oversight (Garcia, 2015). Fully autonomous weapons, which do not yet exist, could operate without such oversight. Such weapons would include artificial intelligence algorithms which would permit them to engage in machine learning, meaning that 'the rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself'* (Matthias, 2004: 177; emphasis added). It is this category that has attracted the most concern. When humans are in or on the loop, they could in principle be held accountable for undesirable outcomes that result from the use of a weapon. This possibility should create incentives for them to carefully consider the law, rules of engagement, and context when using lethal force. But it is not clear who, if anyone, could reasonably be held accountable when fully autonomous weapons produce unwanted outcomes.

## Theory and hypotheses

Responsibility attribution requires that an actor's behavior causally contributes to an outcome (Hewstone, 1991). Autonomous weapons could make this more difficult because of their ability to engage in machine learning. Civilian leaders and military commanders who order the use of such weapons, as well as the engineers who design them, cannot predict precisely how they will operate in an ambiguous environment such as a battlefield. This means they would not exercise sufficient causal control over the weapon to be held accountable for its actions. This difficulty is magnified by the fact that there is no obvious way that autonomous machines can be held accountable for their actions, since they lack intentions, cannot feel physical pain, and cannot experience the negative consequences of public shame and other forms of condemnation (see the useful discussion in Human Rights Watch, 2015). Autonomous weapons would thus make it more difficult to attribute responsibility by increasing the role of those who design the weapon's decision-making algorithms, and by introducing a weapon that could make complex and consequential decisions but could not be held accountable. This would permit human actors to argue that they cannot be held causally responsible, and thus cannot be blamed, for this outcome, an excuse that McGraw terms 'diffusion of responsibility' (McGraw, 1990: 120). Building on this logic, a number of scholars conclude that no one could effectively be held accountable for the outcomes produced by autonomous weapons (Asaro, 2006; Matthias, 2004; Sparrow, 2007).

Since fully autonomous weapons have not been developed, we lack direct empirical evidence that doing so would reduce responsibility attribution to human actors. However, the argument that diffusion of responsibility reduces blame for negative outcomes has been assessed in research on responsibility attribution for economic performance. A key work here is Powell and Whitten (1993). They find little evidence that incumbent governments are punished (rewarded) by voters for poor (good) economic performance. Political contexts that diffuse authority among multiple political actors, such as weak party cohesion, reduce the degree to which incumbents' vote share relates to economic performance. In other words, office-holders who share authority are punished less when things go wrong. These findings, albeit in a different domain, indicate that the introduction of a new entity with the power to make lethal decisions – autonomous weapons – could allow human actors to escape responsibility, have some empirical validity, and suggest the following hypothesis:

> H1: In comparing actions that result in undesirable consequences, those involving more autonomous weapons will result in less responsibility being attributed to civilian leaders and military commanders.

Schulzke (2013) holds that existing mechanisms for responsibility attribution can be extended to the autonomous weapons. His insight is that 'human soldiers are fully autonomous, yet because they act on behalf of their commanders, make decisions within a context that is created by other actors, and rely on intelligence gathered by others, commanders *share* responsibility for soldiers' actions (Schulzke, 2013: 204; emphasis added). When violations of law or policy occur, the contributions of individual soldiers as well as their commanders to the outcome can be determined, and all of those whose behavior contributed to the outcome can be held accountable. The key difference between autonomous weapons and human soldiers is that the former cannot be punished for their actions. But existing accountability mechanisms do not rely solely or entirely on punishing only those who actually use a weapon in combat. This leads Schulzke to conclude that 'to the extent that [autonomous weapons'] actions result from how their

software or hardware is designed, responsibility . . . should lie with the developers who create them. To the extent that their actions are enabled or constrained by civilian and military officials in their chain of command, those officials should share responsibility for the actions of autonomous weapons' (Schulzke, 2013: 204).

Recent work on responsibility attribution leads to two conclusions which both indicate that autonomous weapons might not undermine accountability. First, individuals have a capacity to allocate responsibility among multiple actors. Rudolph (2003), for example, analyzes responsibility attribution to state governments. He finds that, when the same party controls the governor's office and the legislature, the governor is attributed much responsibility. This declines when government is divided. Importantly, though, divided government increases the responsibility attributed to the legislature, indicating that voters are able to shift attributions among political actors as their political influence varies. Second, there is a consistent tendency to attribute much responsibility to actors with greater authority (Kelman and Hamilton, 1989). Experiments reported in Duch et al. (2015) indicate that subjects attribute more responsibility to actors with agenda-setting power. Gomez and Wilson (2001) hold that individuals with less political knowledge blame the most senior political authorities. This means that it might be difficult for political and military leaders to shift blame to other actors. Indeed, Kathleen McGraw, a pioneer in the study of the allocation of blame for political outcomes, finds that attempts to minimize their responsibility by pointing to the diffusion of responsibility are rarely effective (McGraw, 1990, 1991).

This suggests a quite different pattern of responsibility attribution for autonomous weapons:

> H2: In comparing actions that result in undesirable consequences, those involving more autonomous weapons will result in more responsibility being attributed to human actors involved in the design and programming of the weapon.

> H3: In comparing actions that result in undesirable consequences, those involving more autonomous weapons will not result in less responsibility being attributed to civilian leaders and military commanders.

Another line of reasoning challenges the idea that autonomous machines themselves cannot be held responsible. There is evidence that some individuals attribute responsibility to machines and other non-human entities (Epley et al., 2007). Waytz et al. (2010) developed a psychometric measure of such anthropomorphism, the Individual Differences in Anthropomorphism Scale (IDAQ). Subjects scoring higher on this scale place more trust in technological entities to make consequential decisions. If this is the case, anthropomorphism should moderate responsibility attribution:

> H4: In comparing actions that result in undesirable consequences, individuals characterized by higher degrees of anthropomorphism will attribute less responsibility to human actors with greater authority, such as civilian leaders and military commanders, and more to human actors that design autonomous machines as well as to such machines themselves.

## Research design

I designed a survey experiment to assess these hypotheses. Six hundred subjects were recruited from Amazon's Mechanical Turk online labor market. This is not a representative sample; more of the subjects are male and better educated, and they are somewhat more likely to be Caucasian and to identify as a Democrat, than the public at large (see the Appendix for a comparison of the sample's demographic characteristics to the whole population). The use of non-representative samples raises questions about the degree to which the results can be generalized to the population at large. For the present experiment, ensuring such external validity is less important that ensuring internal validity. The weapons described in the treatments are unlikely to be used in combat in the near future. The most relevant population is not today's public, but some future public, and it is difficult to foresee now how the specific characteristics of the weapon system, the nature of the opponent, and so on, could influence this future public's attitudes. The results of the experiment should not be used to predict specific future attitudes but instead make the more modest contribution of testing hypotheses. The Appendix reports the results of regression models that control for subject-level characteristics, including demographic variables as well as attitudes towards the use of force more generally, such as their degree of militant assertiveness (Herrmann et al., 1999). The inclusion of these variables does not alter the pattern of results reported below.

Subjects were randomly assigned to read one of three stories describing a planned air raid on a militant compound in Syria. These treatments stated that the militants controlled air defense systems and that the compound was in an urban area inhabited by civilians. The first (labeled not autonomous) described the raid as conducted by a combat drone operated remotely by a pilot in the United States. The second (semi-autonomous) added that the drone was equipped with sensors that could detect weapons and people in buildings, and that the remote pilot used this information to determine which targets to strike. The third (fully autonomous) stated that the drone was equipped with the same sensors, and that its artificial intelligence computer made decisions about which targets to attack.

All subjects next answered questions measuring their support for the use of force, their expectation that the strike would succeed, and their estimate of the risk that civilians faced. Individual questions were combined to measure each of these concepts (see Appendix for details). All subjects then read another news story stating that the attack described in the treatment took place and resulted in the deaths of dozens of civilians. They were then asked two questions measuring the responsibility and the blame for this outcome borne by the Secretary of Defense, military commanders, the pilot (subjects assigned to the fully autonomous treatment were not asked about the pilot), the engineers who designed the drone's sensors and computer, and the drone's sensors and computers themselves.[1] Subjects answered a question asking if they believed that each of these actors should be investigated for the outcome of the attack. This question assesses the willingness of the participant to support holding the actor accountable. Subjects then completed the individual differences in the anthropomorphism questionnaire (IDAQ). Items dealing with the attribution of human characteristics to technological devices were used to create an IDAQ index for each respondent. The final items on the instrument measured partisanship, age, sex, ethnicity and education.

## Results

Figure 1 displays the degree to which the Secretary of Defense, military commanders, the engineers who designed the weapon, and the weapon itself (described here as its sensors) are responsible for the outcome of civilian casualties for each treatment group.[2] The dependent variable of responsibility attribution is an index combining answers to questions asking if each of these actors is responsible and bears blame for the outcome of the attack; it ranges from zero to one, and higher values indicate more responsibility. The colored dots depict the mean and the bars indicate 95 percent confidence intervals.

Subjects in each treatment group attribute the greatest degree of responsibility to the Secretary of Defense and military commanders and the least to the engineers and the computer and sensors. This is consistent with Hypothesis 3, but not with Hypothesis 1. The responsibility attributed to engineers and the drone's computer increase significantly in the fully autonomous treatment compared to the non-autonomous treatment. This is consistent with Hypothesis 2. Respondents are able to adjust their attributions to include such new actors. This suggests that the development of autonomous weapons would increase the range of actors held responsible for errors.

Figure 2 depicts the mean support for investigating each of these actors; higher values indicate more support for an investigation. On average, subjects exhibit much higher support for the idea of investigating the Secretary of Defense and military commanders than for investigating the engineers and sensors, and this does not vary substantially across the treatments. However, willingness to investigate the engineers and sensors does increase among subjects assigned to the fully autonomous treatment, compared to those assigned to the non-autonomous treatment. This is consistent with Hypothesis 2. Even as the autonomy of the weapon increases, subjects remain more willing to investigate those at the top of the military chain of command. More autonomy also leads to an increased willingness to, at the same time, investigate the role of those responsible for the technical design of the weapon as well as the decisions of the weapon itself, compared to the non-autonomous treatment group.

There are not statistically significant differences in responsibility attribution or support for an investigation between subjects assigned to the non- and the semi-autonomous treatments. On average, subjects do not appear to see great differences between situations where a human is 'in'

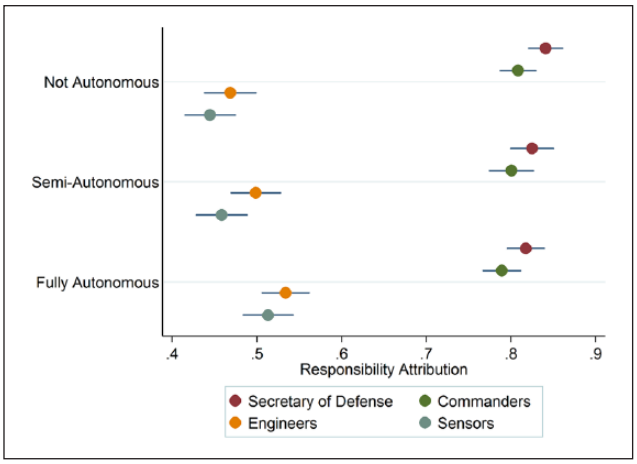


**Figure 1.** Responsibility attribution.

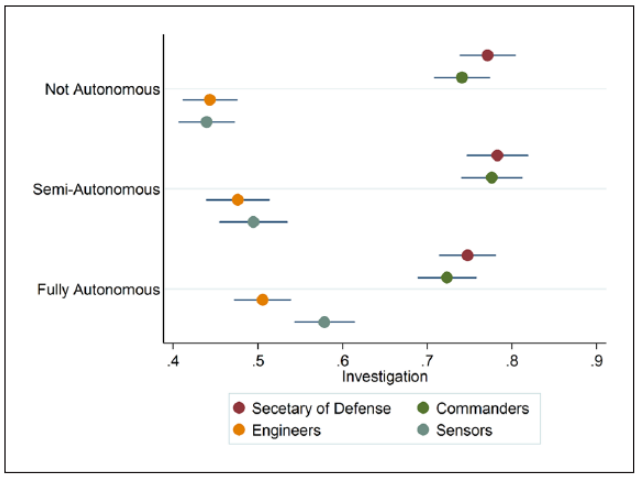


**Figure 2.** Support for investigation.

and 'on' the loop. This indicates that subjects are able to distinguish different degrees of autonomy from each other, and view vesting authority to engage in lethal action in a machine as distinct from situations where machines provide information to human controllers.

These findings are maintained in regression models where responsibility attribution and support for an investigation are the dependent variables, and the independent variables are treatment assignment, militant assertiveness, trust in government, IDAQ, party identification, minority status, age, sex and education (full regression results are reported in the Appendix). Respondents who scored higher on the IDAQ scale attributed less blame to the Secretary of Defense and to the commanders, and attributed more blame to the engineers and sensors. This is consistent with Hypothesis 4. As discussed in the Appendix, these relationships are not contingent on the subject's expectations that the military mission would succeed or would result in civilian casualties.

The regression results suggest other interesting findings. Respondents who score higher on militant assertiveness were less willing to blame the Secretary of Defense or military commanders. This is not surprising, since such individuals view force as useful and legitimate. However, militant assertiveness did not influence the degree of blame attributed to engineers and sensors. The willingness of those who are more assertive to place less blame on civilian and military leaders does not extend to the designers of autonomous weapons or to these weapons themselves. Finally, while the treatments influence responsibility attribution, they do not shape other important attitudes. In particular, support for the use of force does not vary systematically across treatments. Treatment assignment also did not influence the perception that civilian casualties would occur. Both of these findings would seem to merit further investigation into how subjects perceive the capabilities of autonomous weapons.

## Conclusions

Two key conclusions emerge from this experiment. First, autonomous weapons do not decrease the degree to which civilian and military leaders are identified as responsible for negative outcomes. Second, fully autonomous weapons do increase the degree to which those who create and design such a weapon are attributed responsibility. This suggests that the use of fully autonomous weapons may not undermine democratic accountability, as it could create incentives for leaders to carefully oversee such systems and encourage their designers to exercise care in the capabilities they build into such weapons. Of course, even if a robust system of accountability for autonomous weapons were to be established, there may be other strategic and ethical reasons to oppose the introduction and use of this technology, such as concerns about reducing the costs of war, the potential for destabilizing proliferation, or the inability of such weapons to comply with international humanitarian law.

The results of this experiment suggest that fully autonomous weapons would not make it easier for leaders or designers to evade responsibility (McGraw, 1990). Does this mean that such individuals would actually be held accountable for the outcomes such weapons produce? It is difficult to provide a definitive answer to this question. As suggested above, the experiment is designed to assess hypotheses in a controlled environment, and one cannot reliably infer how a future public would react if a use of force like that described in the autonomous treatment were to occur. There is considerable evidence that foreign policy responds to shifts in public opinion, and that leaders seek to minimize military casualties and to highlight battlefield successes in order to influence support for the use of force (see the thorough review in Aldrich et al., 2006). But the effect of public opinion is also mediated by other factors, such as the amount of information individuals have about (potentially covert) military operations, the existence of an elite consensus regarding how to conduct military operations (Kreps, 2010), incentives for legislators to engage in oversight, and the operation of military judicial procedures that operate based on precedent and fact rather than public opinion. This suggests that the degree to which public opinion could influence actual accountability would depend on a range of other independent influences.

These findings have potentially important implications for a range of technologies that rely on artificial intelligence to replace or assist human judgment. Many cyber weapons, for example, are designed to spread across networks and select and attack targets without direct human input. The development of autonomous ground and air vehicles raises important questions about legal and ethical responsibility for accidents. The results reported here suggest that individuals and groups responsible for making the decision to deploy such technologies will continue to accrue a considerable share of the responsibility for undesired outcomes, and that the responsibility and sanctioning of those who create such technologies will increase.

### Acknowledgements

An earlier version of this paper was presented at the workshop 'Man in the Machine', University of St Gallen, June 2015 and the 2015 meeting of the American Political Science Association. For comments and suggestions, I thank Mary Layton Atkinson, Thomas Burri, Justin Conrad, Kate Darling, Graeme Davies, Dirk Lemkuhl, Cherie Maestas, Marcus Schulzke, Markus Wagner and Isabelle Wildhaber.

### Declaration of conflicting interest

None declared.

## Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Notes

1. The president is not included as an actor; see appendix A for a discussion.
2. The average willingness of subjects to blame or investigate the pilot falls between the commanders and the sensors, and does not vary significantly across treatments. Average responses for the pilot are omitted to simplify Figures 1 and 2.

## Supplementary material

The online appendix is available at: http://rap.sagepub.com/content/by/supplemental-data

## References

Aldrich JH, et al. (2006) Foreign policy and the electoral connection. *Annual Review of Political Science* 9(1): 477–502.

Asaro P (2006) What should we want from a robot ethic? *International Review of Information Ethics* 6(12): 9–16.

Department of Defense (2012) Directive: Autonomy in Weapon Systems. Available at: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf (accessed 10 September 2015).

Duch R, Przepiorka W and Stevenson R (2015) Responsibility attribution for collective decision makers. *American Journal of Political Science* 59(2): 372–389.

Epley N, Waytz A and Cacioppo JT (2007) On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114(4): 864–886.

Garcia D (2015) Killer robots: Why the US should lead the ban. *Global Policy Journal* 6(1): 57–63.

Gomez BT and Wilson JM (2001) Political sophistication and economic voting in the American electorate: A theory of heterogeneous attribution. *American Journal of Political Science* 45(4): 899–914.

Herrmann RK, Tetlock PE and Visser PS (1999) Mass public decisions to go to war: A cognitive–interactionist framework. *American Political Science Review* 93(3): 553–573.

Hewstone M (1991) *Causal Attribution: From Cognitive Processes to Collective Beliefs*. New York, NY: Wiley-Blackwell.

Human Rights Watch (2015) Mind the Gap: The Lack of Accountability for Killer Robots. Available at: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots (accessed 10 September 2015).

Kelman HC and Hamilton VL (1989) *Crimes of Obedience: Toward a Social Psychology of Authority and Responsibility*. New Haven, NY: Yale University Press.

Kreps S (2010) Elite consensus as a determinant of alliance cohesion: Why public opinion hardly matters for NATO-led operations in Afghanistan. *Foreign Policy Analysis* 6(3): 191–215.

Matthias A (2004) The responsibility gap in ascribing responsibility for the actions of automata. *Ethics and Information Technology* 6(3): 175–183.

McGraw K (1990) Avoiding blame: An experimental investigation of political excuses and justifications. *British Journal of Political Science* 20(1): 119–131.

McGraw K (1991) Managing blame: An experimental test of the effects of political accounts. *American Political Science Review* 85(4): 1133–1157.

Peffley M (1984) The voter as juror: Attributing responsibility for economic conditions. *Political Behavior* 6(3): 275–294.

Powell GB Jr and Whitten GD (1993) A cross-national analysis of economic voting: Taking account of the political context. *American Journal of Political Science* 37(2): 391–414.

Rudolph TJ (2003) Institutional context and the assignment of political responsibility. *Journal of Politics* 65(1): 190–215.

Schulzke M (2013) Autonomous weapons and distributed responsibility. *Philosophy and Technology* 26(2): 203–219.

Sparrow R (2007) Killer robots. *Journal of Applied Philosophy* 24(1): 62–77.

UN Human Rights Council (2013) Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions. United Nations Human Rights Council, A/HRC/23/4. Available at: http://www.ohchr.org/EN/Issues/Executions/Pages/SRExecutionsIndex.aspx (accessed 10 September 2015).

Wagner M (2014) The dehumanization of international humanitarian law: Legal, ethical, and political implications of autonomous weapon systems. *Vanderbilt Journal of Transnational Law* 47(4): 1371–1424.

Waytz A, Cacioppo J and Epley N (2010) Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science* 5(3): 219–232.